UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Evaluación de métodos automáticos para determinar el umbral óptimo de algoritmos aglomerativos de agrupamiento jerárquico.

Máster Universitario en Investigación e Innovación en TIC (I2-TIC)

Autor: NUÑEZ SALGADO, Alfonso

Tutor: AGUIRRE MAESO, Carlos
Departamento de Ingenieria Informatica

FECHA: Septiembre, 2018

Evaluación de métodos automáticos para determinar el umbral óptimo de algoritmos aglomerativos de agrupamiento jerárquico

Alfonso Núñez Salgado (alfonso.n.s@gmail.com)

7 de septiembre de 2018

Los algoritmos aglomerativos de agrupamiento jerárquicos (AAAJ) son una de las metodologías más extendidas para clasificar y dar sentido a datos de distinto origen. Sin embargo, no existe un único criterio objetivo para determinar la calidad de la clasificación resultante, que permita determinar el punto óptimo para parar el algoritmo de agrupamiento y así obtener de manera automática una clasificación óptima. En este trabajo se proponen nuevas medidas de calidad y criterios de paro y se comparan con las medidas ya existentes consideradas más relevantes a través de datos simulados que permitan evaluar su comportamiento. Además estas medidas se emplearán en la clasificación de proteínas de diferente función en base a su similitud estructural.

Índice

1. Introducción

La gran cantidad de datos biológicos proporcionados por los proyectos de genómica y genómica estructural hacen más necesario que nunca organizar y dar un sentido evolutivo a esta información pues, como ya expuso el genetista Theodosius Dobzhansky, "en biología nada tiene sentido, sino es a la luz de la evolución". Uno de los problemas que más atención ha recibido en este contexto es la clasificación de proteínas en clases estructurales. Al estar la estructura de proteínas mejor conservada que su secuencia, las comparaciones estructurales permiten evidenciar relaciones más remotas que las comparaciones en secuencia y la clasificación estructural de proteínas proporciona una representación exhaustiva del universo de las proteínas conocidas. Tradicionalmente, este problema ha sido abordado por parte de expertos o de forma semi-automática. Sin embargo, tanto por el enorme aumento de la información disponible como por razones conceptuales, es interesante preguntarse si es posible emprender una clasificación estructural de proteínas de manera completamente objetiva y automática, usando distancias estructurales y algoritmos aglomerativos de agrupamiento jerárquico (AAAJ) para esta clasificación. Este programa ha dado resultados muy útiles para la clasificación de proteínas en clases funcionales en base a sus relaciones estructurales [11].

Además, las tradicionales clasificaciones expertas están basadas en dos supuestos implícitos que recientemente se han puesto en duda [18, 27]: (1) Se puede definir de manera objetiva el plegamiento de un grupo de proteínas como clase de equivalencia de estructuras similares, y (2) la clase estructural de una proteína se conserva a lo largo de su evolución. El grupo de investigación con el que realizo el presente trabajo ha contribuido a dicha discusión con la propuesta de cuantificar en qué medida las distancias estructurales entre proteínas cumplen o violan la propiedad transitiva de las clases de equivalencia [22], lo cual tiene implicaciones para la discusión sobre si el espacio estructural de proteínas es discreto o continuo [26, 20]. Se considera el espacio discreto cuando la propiedad transitiva se cumple y continuo cuando se viola, pudiendo pasar con pequeños pasos de un conjunto estructural a otro.

Como resultado, evaluar las violaciones de transitividad en función del umbral del algoritmo de agrupamiento ha permitido ver que el espacio de proteínas es razonablemente discreto para similitudes suficientemente altas, confirmando así el punto de vista tradicional. Sin embargo, también se observa que se hace continuo por similitudes bajas, aunque significativas. De hecho, las clasificaciones estructurales de proteínas más usadas, SCOP y CATH, llegan a agrupar en la misma clase proteínas con baja similitud que violan la propiedad transitiva. Esto hace posible que proteínas de diferente clase se parezcan entre sí significativamente más que proteínas de la misma clase [18, 27]. Desde el punto de vista evolutivo, la fase discreta tendría así una representación en forma de árbol filogenético que se puede atribuir a mecanismos de duplicación génicas seguidos por divergencia estructural. En contraposición, la fase continua tendría como representación una red permitiendo pasar con continuidad de un plegamiento a otro. Esto se puede atribuir a mecanismos evolutivos pluri-parentales en los cuales grandes inserciones o eliminaciones de material genético mediados por recombinación genética.

Un problema abierto de relevancia biológica es como inferir a partir de estas comparaciones una clasificación funcional que permita predecir la función de la proteína a partir de su secuencia y estructura.

1.1. Proteínas, función y reloj molecular

Una proteína es una cadena de aminoácidos codificada en el genoma de un organismo viviente. Con excepción de las proteínas desordenadas, que no se tratarán en este trabajo, la mayoría de las proteínas naturales tienen la propiedad de plegarse formando una estructura tridimensional bien definida, que sin embargo puede variar dependiendo de las interacciones con otras moléculas o cambiando las condiciones ambientales. Cada cadena proteica puede estar formada por uno o más dominios, unidades estructurales relativamente autónomas que se pueden encontrar en distintas proteínas, y que forman la base de las clasificaciones estructurales y funcionales de dominios proteicos. Aunque la definición de la función de una proteína no está fuera de controversia, en este trabajo nos referiremos a función como la definición caracterizada por los términos de *Gene Ontology* que permiten caracterizar la función de un dominio en base a sus interacciones moleculares, a su ubicación en la célula y a la red de procesos biológicos en los cuales interviene.

Los anteriores resultados presentados por el grupo en [21] reflejan que, para altas similitudes tanto en secuencia como en estructura, la divergencia estructural es proporcional a la divergencia en secuencia. Este resultado también se muestra en el trabajo pionero de Chothia y Lesk [9] que, sin embargo, adoptaba una divergencia estructural poco apta a evidenciar esta proporcionalidad. Además, proteínas que cumplen la misma función biológica presentan muy baja diferencia estructural. Estos resultados sugieren que las estructuras de las proteínas de misma función evolucionan según un reloj molecular análogo a lo que describe de manera aproximada la evolución de las secuencias [4]. La hipótesis de un reloj molecular, fue propuesta a finales de los años 60 cuando Emile Zuckerkandl y Linus Pauling [29] mostraron que el número de sustituciones de amino ácidos entre proteínas ortólogas de diferentes especies es proporcional al tiempo que las separa de su ancestro común. De esta manera, estos resultados sugieren que la distancia estructural entre dos dominios de proteínas que cumplen la misma función es proporcional al tiempo que las separa de su ancestro común. Por tanto, que el conjunto de estas distancias estructurales puede ser representado por un árbol filogenético.

Otros resultados del mismo trabajo indican que diferencias estructurales por encima de cierto umbral implican, casi con total certeza, que los dominios comparados tienen distinta función. Además, a partir de ese mismo umbral las divergencias estructurales crecen dramáticamente y de forma no lineal con respecto a la divergencia en secuencia, dejando de describir un proceso evolutivo que cumple el reloj molecular. En este régimen de bajas similitudes estructurales, ocurren grandes inserciones y eliminaciones en las secuencias de las proteínas, lo que implica que el conjunto de distancias se representa mejor con una red que con un árbol. La transición entre estos dos regímenes se puede investigar midiendo las violaciones de la propiedad transitiva.

2. Materiales y métodos

2.1. Particiones y clases de equivalencia

Consideremos un conjunto de N elementos $X = \{x_1, x_2, ..., x_N\}$ que queremos clasificar en K grupos o subconjuntos distintos $C_1, ..., C_k$ de tal forma que un elemento sólo puede pertenecer a un grupo y que cada grupo represente una clase de equivalencia entre los n_k elementos que lo conformen. Dicho de otra manera, buscamos una partición P que verifique:

$$P = \{C_1, C_2, ..., C_k : C_i \cap C_j = \emptyset, \forall C \in P, i \neq j\}$$

$$C_k = \{x_i, x_j, ..., x_h\} \Leftrightarrow [x_i] \sim [x_j] \sim ... \sim [x_h] : \forall i, j \in X$$

El número de particiones posibles de K grupos es 1 cuando K tiene el valor mínimo $K = 1$, por el cual todos los elementos son agrupados y máximo $K = N$ por el cual todos los elementos están separados, y crece de forma exponencial con K para $K \ll N$.

2.2. Propiedades de una partición en clases de equivalencia

Describimos una equivalencia entre dos elementos con el símbolo $x_i \sim x_j$. Una relación de equivalencia posee las propiedades reflexiva, simétrica y transitiva, es decir:

1. $x_i \sim x_i$
2. $x_i \sim x_j \Leftrightarrow x_j \sim x_i$
3. $x_i \sim x_j, x_j \sim x_h \Leftrightarrow x_i \sim x_h$

Además, indicamos el número de elementos en una clase de equivalencia con el símbolo $[x_i] = \{x_j \in X | x_i \sim x_j\}$.

En una red basada en una matriz de distancia, se asocian dos elementos si $d(x_i, x_j) < d_0$. Esta relación posee de manera automática las propiedades de reflexividad y simetría. Sin embargo, la transitividad no siempre se cumple, por lo cual no siempre es posible pasar de manera unívoca de

2.3. Agrupamientos jerárquicos

Buscar el árbol que mejor describe un conjunto de distancias entre elementos es conocido por ser NP-fuerte [23]. Los algoritmos aglomerativos de agrupamiento jerárquico (AAAJ) son una metodología heurística comúnmente utilizadas para generar un árbol a partir de una matriz de distancias. Se trata un tipo de algoritmo de aprendizaje no supervisado que tiene por entrada una matriz con el conjunto de las de distancias entre elementos, y por salida un árbol binario que representa la relación en distancia entre ellos. Esto quiere decir que el error de representación en árbol obtenida está íntimamente ligada a la naturaleza de la matriz de distancias. Esta representación es la mejor posible según un criterio determinado, que es el que diferencia a los distintos AAAJ.

Este tipo de algoritmos empiezan asociando un grupo aun único elemento. De manera iterativa, agrupan los dos grupos con la menor distancia entre sí formando un nuevo grupo y calculan las distancias entre el nuevo grupo y el resto de los grupos. Así sucesivamente hasta que todos los elementos han sido agrupados en un único grupo. Los distintos AAAJ se diferencian entre sí en la forma de calcular la distancia entre grupos. Algunas de las metodologías a las que se hará referencia en este trabajo se muestran en el Cuadro 1.

Como muestra la Figura 1, aplicando un AAAJ hasta un umbral de distancia por el cual todos los elementos están agrupados, se obtiene una partición con un número de subconjuntos entre 1 y N. Particiones que se obtienen con diferentes umbrales están anidadas, de manera que la partición generada con el umbral mayor contiene la partición generada con el umbral más pequeño. En otras palabras, una clasificación jerárquica genera un árbol filogenético y

Metodologías de los AAAJ	Distancia
single-linkage	$D(A,B) := \text{mín}_{a\in A, b\in B}\{d(a,b)\}$
complete-linkage	$D(A,B) := \text{máx}_{a\in A, b\in B}\{d(a,b)\}$
average-linkage	$D(A,B) := \frac{1}{\lvert A\rvert\lvert B\rvert}\sum_{a\in A, b\in B} d(a,b)$
Ward	$D(A,B) := \frac{d(\bar{a},\bar{b})^2}{1/\lvert A\rvert + 1/\lvert B\rvert}$

Cuadro 1: Tabla que muestra algunas de las distintas distancias utilizadas en los AAAJ. Utilizamos $\bar{a}$ y $\bar{b}$ para designar los centros respectivos de cada grupo.

cada punto de corte del árbol corresponde una partición. El objetivo de este trabajo es de determinar el umbral óptimo para agrupar el conjunto de elementos en clases de equivalencia.

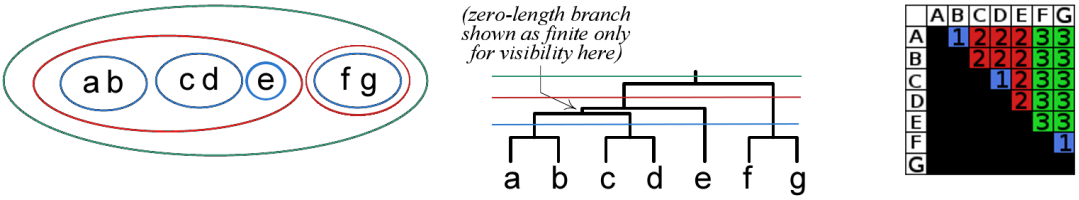

Figura 1: Figura tomada y modificada de [23]. Representación de una clasificación jerárquica. Se muestra los distintos conjuntos de grupos resultantes dependiendo del punto de corte en el árbol filogenético. A la derecha una posible matriz de distancias (distancias por pares de elementos) que puede dar lugar a esta representación.

2.4. Árboles filogenéticos, aditividad y ultrametricidad

Los AAAJ constituyen la técnica más sencilla para reconstruir un árbol filogenético a partir de una matriz de distancia que representa gráficamente sus divergencias evolutivas. Es decir, un árbol binario que representa las relaciones evolutivas entre varias especies u otras entidades que se cree que tienen una ascendencia común. Si la distancia entre dos elementos d_{ij} es proporcional al tiempo evolutivo entre ellas t_{ij}, la matriz de distancias tiene la propiedad de ser ultramétrica. El árbol que describe una matriz de distancias ultramétricas t_{ij} tiene la propiedad de se ser ultramétrico y la distancia que separa dos hojas del árbol pasando por su ancestro común representan tiempo evolutivo entre ellas. Esta distancia tiene la propiedad aditiva, en cuanto la distancia entre dos hojas es igual a la suma de las longitudes de las ramas del árbol que la separan. Como el número de pares de hojas, $N(N-1)/2$, es mucho más grande que el número de ramas, $2N-2$, la propiedad aditiva impone condiciones muy fuertes sobre la matriz de distancias, en particular la propiedad ultramétrica [24] que es mucho más fuerte que la usual desigualdad triangular de las distancias, verificando la siguiente relación :

$$t_{ij} \leq max(t_{ih}, t_{jh}) \forall x_i, x_j, x_h \in X$$

La fiabilidad de un árbol reconstruido con AAAJ depende de manera crítica de que se cumpla la hipótesis de reloj molecular, o sea la distancia d_{ij} entre dos hojas pasando por su antecesor tiene que ser proporcional al tiempo evolutivo entre ellas t_{ij}. Como t_{ij} tiene la propiedad ultramétrica, si el reloj molecular se cumple la matriz M de distancias entre los distintos elementos tiene propiedad de ser ultramétrica. Esto a su vez implica que la matriz de distancia posee la propiedad transitiva de las clases de equivalencia porque, como es fácil ver, si la ultrametricidad se cumple entonces $d_{ik} < d_0$, $d_{jk} < d_0$ y $d_{ij} < d_0$ por cualquier umbral d_0. Sin embargo, ni el reloj molecular, ni la aditividad ni la ultrametricidad se cumplen de manera exacta con los datos reales.

Vemos por tanto la importancia de evaluar la validez de cualquier hipótesis que asuma que en los datos existe una estructura compatible con un árbol y, en particular, si dicha hipótesis se sostiene para todas las entidades biológicas sometidas al análisis o solamente un subconjunto de las mismas. Es decir, nos preguntamos si, a lo largo de la construcción de un árbol, es posible determinar de manera objetiva cuándo los datos dejan de ser consistentes con una estructura de árbol y debemos parar el AAAJ.

2.5. Criterios de calidad de la partición y puntos de paro

Se considera que el criterio de paro de un AAAJ debe generar la partición óptima. Indicando el corte en el en el punto más alto árbol que garantice que las ramas que quedan por debajo (figura 1) son grupos de elementos que pertenecen a una misma clase de equivalencia. Aunque no existe un criterio de paro universal, actualmente sí existen más de una treintena de índices para medir la calidad de una partición [12] [8]. Sin embargo, por norma general, no existen índices que permitan encontrar un punto de paro objetivo, o formal, en todo el rango de las N posibles particiones generadas por el algoritmo. Por esta razón, la mayoría de técnicas actuales hacen uso de conocimiento *a priori*, buscando la partición óptima en un rango determinado muy inferior al número total de elementos, y establecen como punto de paro el valor extremo (mínimo o máximo absoluto) del índice utilizado. Todos estos índices se basan, de una forma u otra, en relacionar las distancias entre los elementos del mismo grupo d^{intra} con las distancias de los elementos entre distintos grupos d^{inter}. El objetivo de este trabajo pretende encontrar la partición óptima sin ningún conocimiento *a priori*, abarcando todas particiones que se pueden generar a lo largo del AAAJ, se han desarrollado nuevos índices de calidad a los que se puede aplicar criterios objetivo de paro teniendo en cuenta las siguientes propiedades deseables.

- Estar definido para cualquier partición.
- Tener el mismo valor cuando todos los elementos están agrupados en un único grupo y cuando existe un único elemento por grupo.
- Encuentra la partición óptima donde existe un mínimo o máximo absoluto del índice respecto a todas las particiones generadas por el algoritmo de agrupamiento.

Hemos seleccionado de la literatura una serie de índices de calidad que nos han parecidos relevantes, ya sea por su comparada efectividad [19], por su relevancia con respecto a otros

índices o por similitud a los nuevos índices desarrollados. Aunque la mayoría de índices seleccionados no tienen las propiedades deseadas, en algun caso se han podido realizar ciertas modificaciones para obtenerlas.

2.5.1. Violación de transitividad : Una nueva medida

La violación de la tercera propiedad de una clase de equivalencia, la transitividad, puede medirse de varias maneras. Una de ellas es midiendo la violación de transitividad que se genera al juntar dos grupos en cada paso del AAAJ. Este es el caso de la medida de violación de transitividad (TV^{old}) publicada por grupo donde se está realizando este trabajo [22]. Si consideramos A,B y C tres grupos pertenecientes a la partición generada en el paso t, la medida vendría definida por:

$$TV^{\text{old}}(ABC) = \frac{D_{AC} - D_{BC}}{D_{AC} - D_{AB}}, D_{AC} \geq D_{BC} \geq D_{AB}$$

Donde, siendo w_c un peso proporcional al número de elementos del conjunto C, el error cometido al juntar dos subconjuntos en relación a la propiedad transitiva sería:

$$TV^{\text{old}}(A + B \rightarrow AB) = \sum_{C \neq A,B} w_c TV^{\text{old}}(ABC)$$

Esta medida cuantifica la violación de la transitividad que se genera al juntar dos grupos. Teniendo un valor comprendido en el rango [0,1] donde 1 implicaría la máxima violación y 0 que no existe tal violación. El problema de la formulación de esta medida es que no es una función de estado que se puede calcular de manera unívoca para cada partición, sino que depende de los pasos realizados por el AAAJ. Además, aunque se formulara como tal, requiere de un número elevado de operaciones ya que es necesario tomar en cuenta todos los posibles tripletes de elementos entre conjuntos.

La violación de la transitividad se produce cuando un elemento de un grupo está más próximos a otro elemento de distinto grupo que a elementos del suyo mismo. Está directamente relacionada con el solapamiento de las distribuciones de las distancias d_c^{intra} y d_c^{inter}. Por tanto, otra manera de medir la violación de la transitividad es cuantificar este solapamiento.

Para este trabajo se ha desarrollado una nueva definición de violación de transitividad que, a diferencia de la medida anterior, computa la violación de la transitividad de forma global para toda la partición P. Podemos decir que si la distancia entre los elementos pertenecientes a un subconjunto (d_c^{intra}) supera cualquiera de sus distancias a elementos de otros subconjuntos (d_c^{inter}) existe una violación de transitividad que puede ser cuantificada de la siguiente manera, viéndose representada en la Figura 2 :

$$TV = \sum_{k=1}^{K} \sum_{\substack{i<j \\ i,j \in C_k}}^{N_k} \vartheta(d_{ij} - d_0)(d_{ij} - d_0) + \sum_{k<m}^{K} \sum_{\substack{i \in C_k \\ j \in C_m}}^{N_k, N_m} \vartheta(d_0 - d_{ij})(d_0 - d_{ij}) \tag{1}$$

En esta fórmula, $\vartheta(x)$ designa la función de Heaviside, que tiene valor 0 cuando $x < 0$ y 1 cuando $x \geq 0$. Y el parámetro d_0, en este caso, designa el umbral donde solapan ambas distribuciones.

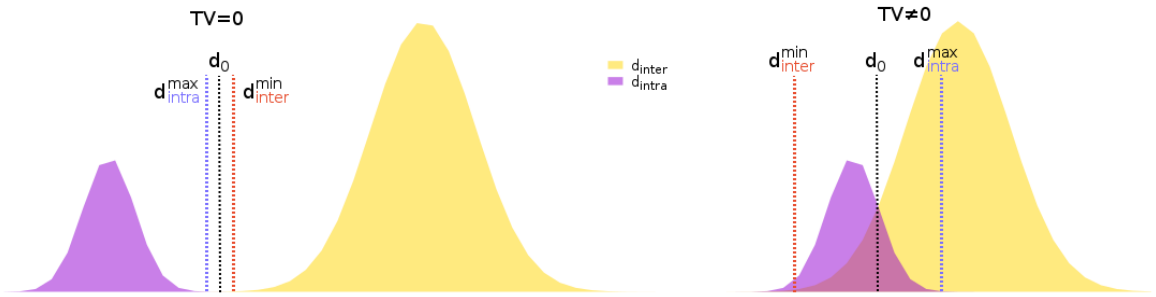

Figura 2: Representación esquemática las distribuciones de d_c^{intra} (morado) y d_c^{inter} (amarillo) en dos casos distintos. En el caso de la izquierda no existe solapamiento por lo que $TV = 0$. En el caso de la derecha existe solapamiento por lo que elementos de un mimo grupo son más cercanos a elementos de otro grupo que a los del propio. En este último caso $TV \neq 0$.

Por tanto, para cada partición P, hay que calcular el umbral d_0 comprendido entre $\max(d_c^{\text{inter}})$ y $\min(d_c^{\text{inter}})$ que minimiza esta medida. Para simplificar nuestros cálculos, hemos utilizado un valor aproximado que es no en muy distinto del óptimo en los ejemplos que hemos examinado.

$$d_0 = \frac{max(d^{\text{intra}}) + min(d^{\text{inter}})}{2} \tag{2}$$

Esta medida puede ser interpretada como una distancia o disimilitud entre las distribuciones d_c^{inter} y d_c^{intra}, está definida entre $[0, +\infty)$ y es dependiente de la escala de las distancias.

Sin embargo, esta medida toma su valor mínimo para las particiones "triviales" $K = 1$ y $K = N$ y por lo tanto no se puede usar como criterio de paro, ya que no vamos a encontrar un máximo (o mínimo) en una partición intermedia. En este trabajo, la usaremos para cuantificar la calidad de la partición en el punto de paro.

2.5.2. Coeficiente de Bhattacharyya

Podemos decir que cuando las distribuciones de las distancias d^{intra} y d^{inter} solapan existe violación de la transitividad dentro de los elementos que lo conforman. Por tanto, toda medida que disimilitud entre funciones de densidad de probabilidad (PDF) que tenga en cuenta el solapamiento es sensible de ser usada como medida de violación de transitividad. Una de estas medidas viene dada por el coeficiente por Bhattacharyya en 1943 (BC)[2]. Siendo p y q dos funciones de probabilidad definidas en el mismo dominio X, su formulación sería la siguiente :

$$BC(p, q) = \sum_{x \in X} \sqrt{p(x)q(x)}$$

Esta medida está comprendida entre [0, 1] y tiene una interpretación estadística/probabilística, estimando la proporción de solapamiento entre ambas distribuciones. Además esta medida acota la probabilidad de frustración de la clasificación en términos bayesianos[6]. Sin embargo, al igual que todas las medidas basadas en PDF, es computacionalmente costosa ya

que implica calcular los histogramas de ambas distribuciones y determinar el ancho del *bin* de estos histogramas no es una tarea trivial: (a) tiene que ser el mismo para ambos histogramas y (b) no puede haber ningún *bin* vacío.

Este índice no es objetivo de estudio como índice de paro pero ha sido utilizado para comparar los resultados obtenidos en el índice TV. También se han tomado en cuenta distancias que derivan de este mimo coeficiente. La distancia de Bhattacharyya ($-\log(BC)$) y la de Hellinger ($\sqrt{1-BC}$), que a diferencia de la anterior es una métrica acotada y garantiza la desigualdad triangular. Esta propiedad es deseable porque nos permite comparar de forma inequívoca tres o más grupos de elementos en base al error cometido al ser agrupados en términos transitividad. Estas métricas no se presentarán en la comparativa más adelante ya que no muestran diferencias significativas.

2.5.3. Índice de Dunn

El índice de Dunn (DI), criterio desarrollado por J. C. Dunn en 1974 establece una relación relativamente simple entre d^{intra} y d^{inter} [13]:

$$DI = \frac{\min(D_{C,C'})}{\max(d_c)}$$

A partir de esta relación se han establecido distintos criterios que más tienen que ver con la definición de las distancias que con la relación en sí. Para poder satisfacer las propiedades mencionadas anteriormente, y sólo para este índice, hemos elegido las siguientes definiciones de distancia :

$$\begin{aligned} d_c &= \frac{1}{n_c(n_c-1)} \sum_{\substack{i>j \\ i,j\in C'}} d_{ij} \\ D_{C,C'} &= \frac{1}{n_c n_{c'}} \sum_{\substack{i\in C \\ j\in C'}} d_{ij} \end{aligned} \tag{3}$$

Dada esta definición el valor del índice no pude ser inferior a 0 ya que $d_c \leq D_{C,C'}$, pero no tiene cota superior. Cuanto mayor es el valor del índice mejor es la partición encontrada y por tanto el punto de paro se encuentra en el máximo absoluto.

2.5.4. Silhouettes

El índice Silhouettes desarrollado por Peter J. Rousseeuw en 1986 [25] establece una medida para cada elemento. Definiendo $d_{\text{in}}(i)$ como la distancia media del elemento al resto de elementos de su propio grupo, y $d_{\text{out}}(i)$ como la distancia media del elemento al resto de elementos que no pertenecen a su grupo:

$$d_{\text{in}}(i) = \frac{1}{n_c - 1} \sum_{\substack{j=1 \\ j\neq i}}^{n_c-1} d_{ij}, \forall i, j \in C$$

$$d_{\text{out}}(i) = \frac{1}{N - n_c} \sum_{\substack{j=1 \\ j \in C'}}^{N-n_c} d_{ij}, \forall i \in C, j \notin C$$

la medida vendría formulada por:

$$s(i) = \frac{d_{\text{out}}(i) - d_{\text{in}}(i)}{\text{máx}\{d_{\text{in}}(i), d_{\text{out}}(i)\}}$$

Y la calidad del partición se cuantifica con la media de $s(i)$ entre todos los elementos:

$$SI = \frac{1}{N} \sum_{i=0}^{N} s(i) \tag{4}$$

Esta medida está comprendida entre 1 y -1 y siempre que $d_{\text{in}}(i) < d_{\text{out}}(i)$ el valor de $s(i)$ será positivo, por tanto buscaremos el valor máximo de SI. Sin embargo SI no está definida para particiones en las que las exista un único elemento en un grupo, o que todos los elementos estén en un único grupo. Para poder computar el índice a lo largo del algoritmo de agrupamiento establecemos dos definiciones :

- $d_{\text{in}}(i) = 0$ cuando $|C_k| = 1$, $i \in C_k$
- $SI = 0$ cuando k=1. Esta definición es razonable ya que en este caso ni $d_{\text{in}}(i)$ ni $d_{\text{out}}(i)$ están definidas para esta partición y nos permite obtener las propiedades deseadas.

Para garantizar las propiedades deseadas para la función de evaluación del punto de paro queremos que el valor del índice sea el mismo cuando K=1 y cuando K=N. Ya que cuando K aumenta SI tiende a 1 podremos obtener este resultado restando al índice original la pendiente $((K-1)/(N-1))$, obteniendo entonces la siguiente función de evaluación de punto de paro:

$$SI^{\text{mod}} = \frac{1}{N} \sum_{i=0}^{N} s(i) - \frac{K-1}{N-1} \tag{5}$$

Esta definición permite definir el punto de paro en el máximo absoluto. De aquí en adelante nos referiremos al índice SI^{mod} como SI.

2.5.5. Criterio de Calinski-Harabasz

El criterio desarrollado por T. Caliński y J. Harabasz en 1974,[5] también conocido como *Variation Ratio Criteria* (VRC) es uno de los criterios más robustos [19] siempre que se busque en un rango pequeño de K. Este índice se basa en la propiedad de que la matriz de distancias elevadas al cuadrado es igual al la suma de la totalidad de las distancias al cuadrado de los elementos a centro de masa del grupo a a los que están asociados (TGSS, *Total Group Sparse Square matrix*). De forma análoga, define el conjunto de distancias al cuadrado de los elementos del mismo grupo como WGSS (*Within Group*) y las distancias al cuadrado entre grupos como BGSS (*Between Groups*). Estableciendo las siguientes relaciones:

$$CH = \frac{BGSS/(K-1)}{WGSS/(N-K)} \tag{6}$$

Donde

$$TGSS = WGSS + BGSS$$

$$TGSS = \sum_{1}^{N} ||x_i - \mu||^2 = \frac{1}{N} \sum_{\substack{i=1 \\ i>j}}^{N} d_{ij}^2$$

$$WGSS = \sum_{1}^{k} \sum_{1}^{n_c} ||x_i - \mu_c||^2 = \sum_{1}^{K} \frac{1}{n_c} \sum_{\substack{i>j \\ j,i \in C^k}} d_{ij}^2$$

Este criterio tiene una cota inferior $[0, +\infty)$, no está definido para K=1 ni para K=N y tiende a crecer con K. Además es muy sensible a la cantidad de elementos a clasificar, alcanzando a magnitudes tales que que su representación gráfica puede carecer de sentido. Aunque este índice presenta multitud de máximos y mínimos locales, el primer máximo local es el más significativo, siendo este punto el utilizado como punto de paro.

2.6. Datos simulados

Nuestro objetivo es comparar el comportamiento de los criterios de paro y las medidas de calidad en condiciones controladas donde se pueda generar una clasificación con creciente frustración. Es decir generar una matriz de distancias donde no cumple la propiedad transitiva, por lo cual las clases de equivalencia no se dan de manera exacta sino que hay que encontrar un compromiso entre separar elementos parecidos y juntar elementos diferentes. En estas condiciones, queremos comparar la clasificación de referencia con las clasificaciones obtenidas con los varios criterios de paro. Además, queremos garantizar que la disimilitud entre los elementos generados es una distancia que verifica la propiedad triangular, por lo que, por conveniencia, generamos datos como puntos en un espacio euclídeo y calculamos su matriz de distancias.

En primer lugar generamos N_C puntos equidistantes (K-simplex) al rededor de una hiperesfera de radio *Radi* en en espacio N-dimensional (con $N_c - 1$ dimensiones), a los que llamaremos centros. A partir de estos centros se generan de forma aleatoria n_c puntos siguiendo una distribución $\mathcal{N}(N_c, \sigma^2)$. Finalmente se reescalan los puntos para que la distancia máxima se inferior o igual a 1. Como se muestra en la figura 3,variando los parámetros σ^2 y *Radi* podemos solapar la distribuciones de puntos y generar clasificaciones frustradas. Como en este ejercicio la función de estos parámetros es equivalente (solapar distribuciones), siempre que en un experimento uno varíe el otro quedará fijo a un valor determinado.

2.7. Frustración de la partición simulada

Se ha establecido una métrica de control F que tiene por objetivo medir la frustración esperada en la clasificación. F se ha definido como la proporción de puntos que están más cerca de un centro distinto al que se usó para ser generados. N_F es el número total de puntos que cumple esta condición y puede ser calculado en el momento en el se generan los puntos.

$$F = \frac{N_F}{N}$$

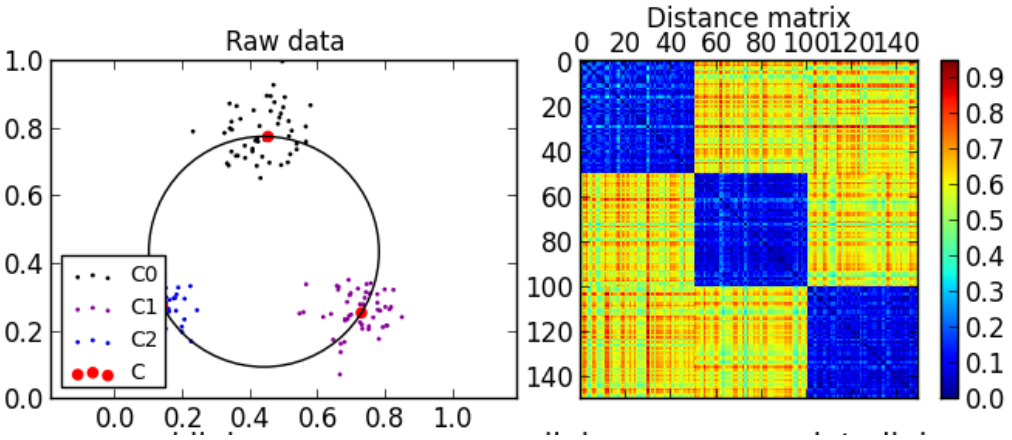

Figura 3: Representación gráfica del modelo utilizado (a la izquierda) y la matriz de distancias generada (al aderecha)

2.8. Comparación con una partición de referencia

Para comparar la clasificación obtenida con la original generada se han empleado los siguientes índices:

1. La diferencia entre número de particiones original y el encontrado. Índice simple que nos permitirá ver si el punto de paro asociado a cada criterio tiende a agrupar o por el contrario a dividir en grupos más pequeños los elementos

2. Rand Index (RI) : es uno de los criterios más comunes que establece una similitud entre particiones, definiéndose entre el rango $[0; 1]$ donde 1 implica que ambas particiones son idénticas y 0 que son completamente distintas.

3. Adjusted Rand Index (ARI): Es lanormalización de Rand Index (RI) por su esperanza. RI no toma en cuenta el número de grupos que existe en una partición así dos particiones distintas que se diferencian simplemente en que un grupo de la primera es subdividido en dos grupos distintos en las segunda tendrán un valor $RI = 1$ y un valor $ARI < 1$.[[28]]

4. V-measure (VI): Equivalente a la información mutua normalizada entre ambas particiones que puede ser interpretada como una medida de disimilitud entre dos particiones. Está definida entre $[0; +\infty)$ donde 0 implica que las dos particiones son idénticas.

2.9. Clasificación funcional de superfamilias de proteínas

Si bien las proteínas muy distantes estructuralmente tiene distinta función y las que tienen misma función son cercanas, es interesante comprobar si todas aquellas que son suficientemente cercanas tienen la misma función. Si esta condición se cumple, el uso de AAAJ para predecir la función es una metodología suceptible de ser utilizada.

Con este objetivo se han elegido tres de los cinco conjuntos de dominios (unidades estructurales) de proteínas publicados por el grupo en [21]. Cada conjunto esta formado por un conjunto de dominios que tienen menos de un 40 % de identidad en secuencia (semejanza entre ellos) y que coinciden en la forma en que han sido clasificados entre CATH y SCOP,

dos de las clasificaciones estructurales de proteínas más usadas. Estos conjuntos de dominios pertenecen a tres superfamilias: Aldolasas (TIM barrel fold), P-loop NTPase y NADP-binding Rossman-fold. De todos estos conjuntos se han eliminado las estructuras que contengan dominios que contengan varias cadenas de polipéptidos ya que la asignación de función ha sido problemática, y se ha elegido un único representante de aquellos dominios con una identidad, tanto en secuencia como en estructura, superior a 0,98.

La asignación de función a estos dominios se ha hecho a través de los términos de *Gene Ontology* (GO) que permiten caracterizar la función de cada dominio en base a sus interacciones moleculares, a su ubicación y a la red de interacciones proteicas donde actúa. La asignación de la función de se ha de tal manera que dos dominios tiene la misma función si todos los términos de GO que la describen son idénticos.

Si consideramos una función binaria que devuelve un valor positivo cuando se juntan dos dominios con la misma función a lo largo de AAAJ y negativo cuando tienen diferente función, se puede caracterizar la capacidad del algoritmo de agrupar dominios con la misma función mediante una curva ROC (Receiver Operating Characteristic), relación ordenada entre el ratio de verdaderos positivos y el ratio de falsos positivos. Los resultados de esta caracterización se muestran en es misma publicación, obteniendo para cada conjunto de datos valores de área bajo la curva muy cercanos a uno (AUC). Por tanto existen indicios de que la medida utilizada como disimilitud estructural junto con la metodología utilizada puede tener la capacidad de predecir la función.

El número de dominios y el número de funciones a signados en cada domino en los conjuntos de datos es distinto y se ha resumido en el Cuadro 3.

Superfamily	AUC
Aldolases	0.980
P-loop	0.973
NADP	0.812

Cuadro 2: Área bajo la curva ROC resultante de la clasifición funcional de los dominios publicados en [21].

Datos	Dominios	Número de funciones
Aldolasas	272	21
NADP	676	50
P-loop	533	53

Cuadro 3: Número de dominios y grupos funcionales en los conjuntos de datos.

3. Resultados

3.1. Cluster Information Criteria, un nuevo criterio de paro

En este apartado introducimos un nuevo criterio de paro inspirado en teoría de la información. Supongamos que los puntos i que pertenecen al grupo c, o sea con $C_i = c$, son generados de forma independiente por una distribución Gaussiana con parámetros p_c, μ_c y σ_c

$$p(x_i|C_i = c) = \frac{p_c}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-(x_i-\mu_c)}{(2\sigma_c^2)}} \tag{7}$$

Podemos estimar los parámetros de esta distribución maximizando la probabilidad de los datos observados dados los parámetros, usando como única información de entrada una matriz

de distancias:

$$\sigma_c^2 = \frac{\sum_{ij} d_{ij}}{n_c^2} \quad p_c = \frac{n_c}{N} \tag{8}$$

donde n_c es el número de elementos en la clase c. La *log-likelihood* de los datos se puede calcular como:

$$\ln(\mathcal{L}) = \sum_i \ln(\mathcal{P}(x_i|C_i = c)) = n_c \cdot \ln(p_c) - n_c \cdot \ln(\sigma_c^2) + n_c \cdot \text{const}$$

Si tomamos un elemento del grupo como representante y lo usamos para definir el parámetro μ, lo tenemos que quitar del cálculo de la probabilidad: Grupos con un solo elemento no dan ninguna información porque $x_i = \mu_c$ y $\ln(P) = 0$. Obtenemos así la fórmula

$$\ln(\mathcal{L}) = (n_c - 1)ln(p_c) - (n_c - 1)\ln(\sigma_c^2) + (n_c - 1) \cdot \text{const}$$

Sin embargo, cuanto mayor es el número de grupos K, mayor es el número de parámetros y más alta la probabilidad, obteniendo una probabilidad de 1 cuando $K = N$. Por lo tanto, tenemos que penalizar los parámetro del modelo, cuyo número es $2K$, de forma similar a como lo hace el criterio de información de Akaike (AIC), obteniendo un índice que denominamos CIC (Cluster Information Criteria) definido por:

$$\text{CIC} = \frac{1}{2}AIC = \text{Npar} - \ln(\mathcal{L}) \tag{9}$$

$$\text{CIC} = (K-1)\mathcal{P}(p) + (K-1)\mathcal{P}(\sigma_c^2) - \sum_{C=1}^{K} n_c \ln(p_c) + \sum_{C=1}^{K} \frac{n_c}{2} \ln(\sigma_c) + \sum_{C=1}^{K} n_c \cdot \text{const} \tag{10}$$

Este índice es ≤ 0 y toma el valor 0 en correspondencia con las particiones triviales $K = 1$ y $K = N$, por lo cual posee un mínimo absoluto en correspondencia de una partición no trivial que permite determinar un punto de paro.

3.2. Evaluación con datos simulados

3.2.1. Violación de Transitividad y solapamientos entre distancias *inter* e *intra* cluster

En primer lugar comprobamos en al figura 4 que obtenemos los resultados esperados, que todas las métricas aumentan en magnitud a medida que σ^2 aumenta, empezando a crecer desde que $\sigma^2 > 0{,}2$. Y mientras F_{ratio} como TV lo hacen de de una forma casi lineal BC lo hace de una forma logarítmica.

La métrica de control F cambia significativamente en su pendiente a medida que el numero de centros aumenta, y por tanto el número total de elementos, mostrando que la proporción de elementos en intersección proporcionalmente. Este efecto se ve de igual manera en BC disimulado por su crecimiento logarítmico, sin embargo en criterio TV no se ve especialmente afectado por este fenómeno a partir de $N_c = 3$, mostrando que este criterio es mucho más

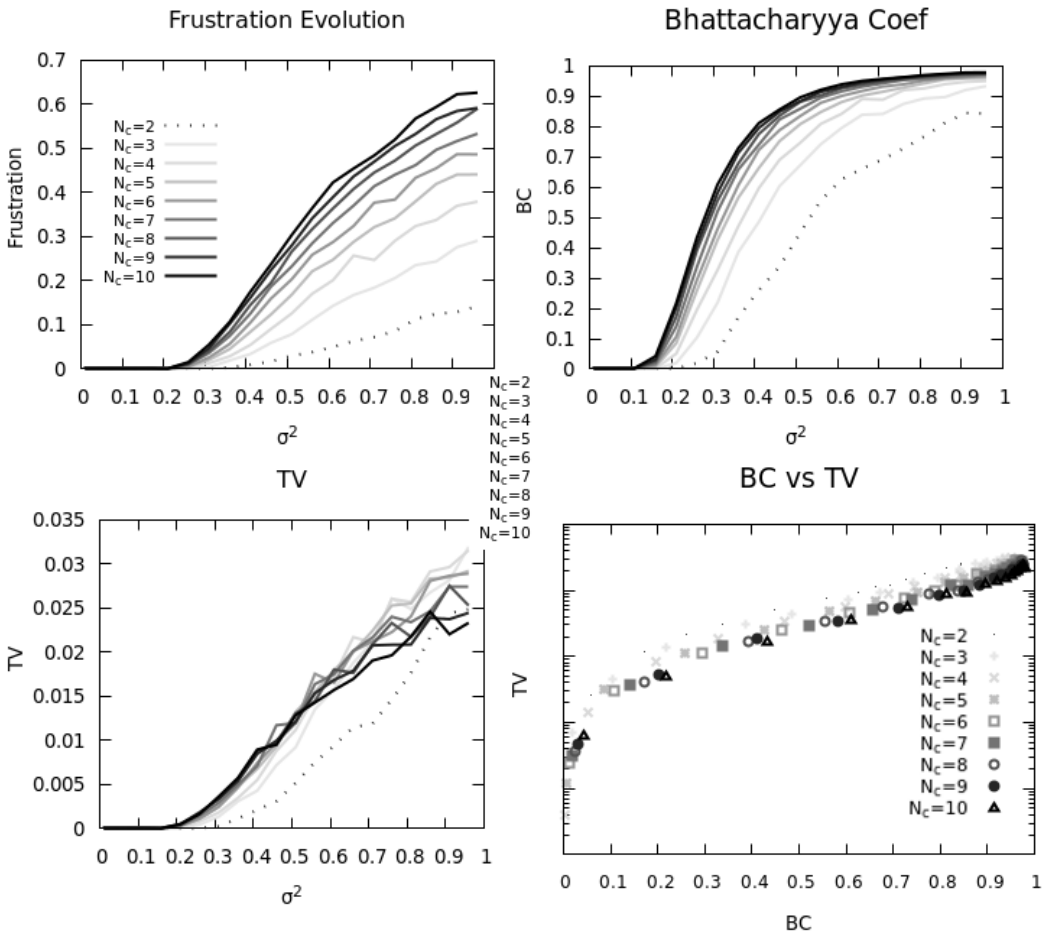

Figura 4: Figura en la que se representa la frustración de la clasificación simulada en función de los parámetros (arriba a la izquierda), el solapamiento de d^{intra} y d^{inter} de la partición original (arriba a la derecha) y la violación de transitividad (abajo izquierda). Ademas se muestra la correlación entre las dos medidas de calidad BC y TV, donde TV tiene escala logarítmica (abajo a la derecha). En estas gráficas se muestran los resultados medios obtenidos con 30 generaciones de datos independientes por cada parámetro barrido, dejando fijo tanto el radio ($R = 1$) como en número de elementos por centro ($n_c = 30$), y variando tanto σ^2 como número de centros

robusto con respecto al número de elementos permitiendo comparar la violación de transitividad entre espacios donde varíen el número de grupos o clases. Esta es una propiedad deseable para una medida de calidad.

Si bien la pendiente de crecimiento de del esta métrica se muestra independiente del número de centros generados muestra un crecimiento en dientes de sierra para valores de σ^2 altos, seguramente debido a la estimación d_0 (ecuación 2). Sin embargo esta estimación no parece afectar a situaciones donde la violación de transitividad es baja y el número de elementos que frustren la clasificación no supere el 30 % del total de los elementos. Además

observamos que existe una correlación entre $\log(TV)$ y BC, por lo que deducimos que TV tiene una relación directa con la superficie de solapamiento entre dos distribuciones y relaciona en términos de distancias y similitudes dos distribuciones gaussianas.

3.2.2. Caracterización de los criterios de paro

A diferencia del apartado anterior los conjuntos de datos generados van a ser clasificados por un algoritmos de agrupamiento jerárquico y su clasificación va a ser comparada con los grupos originales.

Aunque se han aplicado distintos algoritmos de agrupamiento (*Average linkage, Single Linkage, Complete Linkage, Median linkage, Weighted Linkage, Centroid Linkage* y *Ward*), en todos los casos mostrados se ha empleado el algoritmo de Ward, porque al garantizar la mínima varianza de distancias entre los elementos de una rama permite que conjunto de elementos sea más compacto.

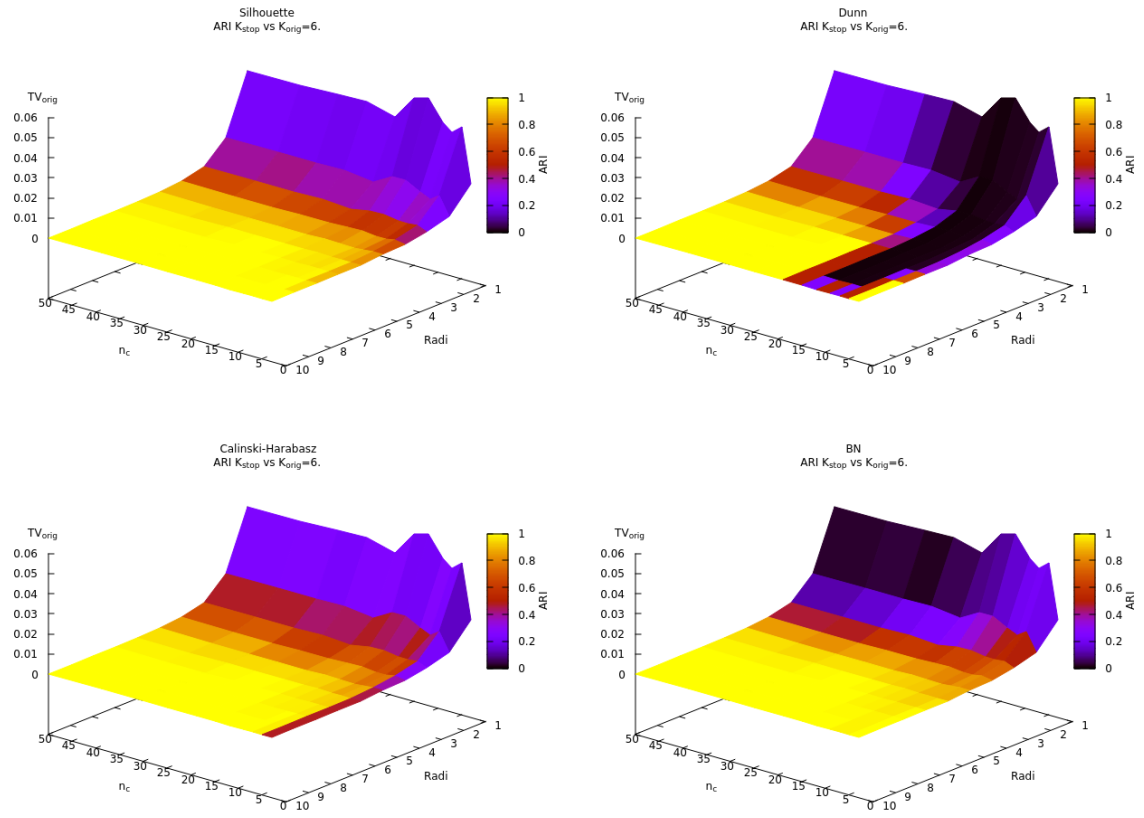

Figura 5: Figura que muestra el la calidad clasificación resultante con respecto la original (ARI) al aplicar los distintos criterios de paro automáticos. Los datos se han generado siguiendo los criterios especificados en el experimento k-simplex fijo el parámetro $\sigma^2(=0.1)$ y el número de centros, parámetro $N_c(=6)$

Observamos tanto en las figuras 5 y 6 que la violación de transitividad es consistente a lo largo del barrido de parámetros siendo más irregular cuando el número de le puntos pos

centro es más pequeño. Esto puede ser debido por un lado a la estimación aproximada de d_0 y por otro a que el número de repeticiones de la prueba no es muy grande.

Los criterios de paro SI y CH recuperan el número de centros originales incluso en situaciones donde la violación de transitividad es alta, aunque en estos casos como es de esperar la partición encontrada no corresponde con la original. Por ejemplo se obtienen ARI inferior 0,2 cuando la TV es superior a 0.2. Ambos criterios de paro SI y CH tienden a recuperar un número de centros que establecen sus respectivos puntos de paro (K_{stop}) inferior al original (K_{orig}) cuando la violación de la transitividad es elevada, es decir que tienden un sesgo para agrupar más que en la partición original.

El número de centros encontrado con el criterio de paro de DI es llamativamente elevado cuando el número de puntos por centro está por debajo de 20, además la partición dista mucho de la original incluso para situaciones donde no existe violación de transitividad. Además el criterio de Dunn parece ser especialmente sensible a la cantidad de datos tendiendo a agrupar más cuando el número de datos de entrada aumenta.

El criterio desarrollado en este trabajo (CIC) recupera el número de centros y la partición original siempre que la violación de transitividad sea baja o nula, sin embargo aumenta drásticamente en el momento en que existe solapamiento entre las distribuciones de puntos generados Esto se debe a que el índice empieza a crecer debido a que los grupos que se conforman a partir de cierto punto (mínimo en el criterio de paro) no siguen una distribución gaussiana.

A la luz de estos resultados podemos decir que los puntos de paro obtenidos con los criterios CH y SI consiguen recuperar mejor la información original , mientras que CIC encuentra puntos de paro donde el número de grupos K_{stop} es muy superior a K_{orig} en situaciones en las cuales la violación de transitividad de los datos originales es grande, o sea separa más que la partición original, garantizando así que los elementos de grupos encontrados pertenezcan a la misma clase de de equivalencia, pero no consigue separar grupos que son distintos en las simulaciones.

3.3. Clasificación de proteínas en clases funcionales

En primer lugar realizamos la clasificación de los conjuntos de datos descritos en Materiales y métodos con distintos algoritmos de agrupamiento jerárquico y medimos la calidad de los particiones generadas a lo largo del algoritmo con respecto a la clasificación de la función de la proteína esperada (Figura 7). En segundo lugar hallamos los puntos de paro correspondientes a los índices descritos en secciones anteriores utilizando Ward como AAAJ.

3.3.1. Agrupamiento jerárquico de los datos

Los resultados de la Figura 7 nos muestran que en el caso de las Aldolasas sí conseguimos una clasificación satisfactoria a través de distintos algoritmos de agrupamiento, llegando a valores $ARI > 0{,}90$ y $VI < 0{,}29$ en todos los casos. Además el número optimo de grupos encontrado (k_{VI} y k_{ARI}) dista por poco del número de grupos de funciones ($K_{\text{Aldolase}} = 21$) dependiendo del algoritmo de agrupamiento usado.

Este resultado muestra que estos datos contienen conjuntos de distancias entre elementos con una clara forma de árbol y valida en cierta medida la metodología empleada para el cálculo de las distancias entre los dominios de proteínas, aunque no la exime de todo error

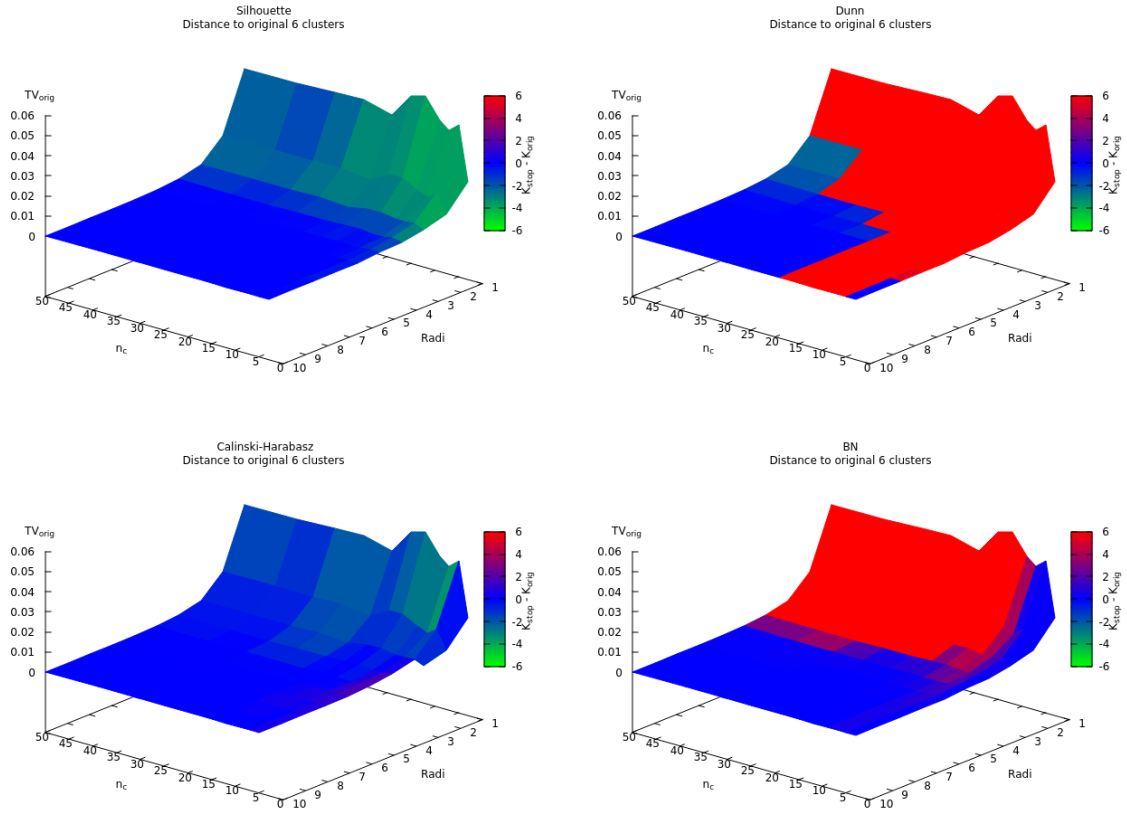

Figura 6: Figura que muestra la diferencia entre el número de grupos encontrados K_{stop} con respecto la original $K_{\text{orig}} = N_c$ al aplicar los distintos criterios de paro automáticos. Los datos se han generado siguiendo los criterios especificados en el experimento k-simplex fijo el parámetro σ^2(=0.1) y el número de centros, parámetro N_c(=6)

ya que en ninguno de los algoritmos de agrupamiento utilizados se consigue encontrar una clasificación optima ($ARI = 1$ o $VI = 0$).

En este sentido hubiera sido conveniente utilizar el índice de Rand (RI) no normalizado, ya que mostraría en que punto del algoritmo empiezan a aparecer errores en la clasificación. Esto aportaría información de mayor relevancia ya tanto ARI como VI muestran comportamientos similares y permitiría comparar de forma más eficaz los distintos algoritmos de agrupamiento, ya que tendría un valor de 1 en $k = N$ e iría decreciendo a medida que el AAAJ junte grupos que no tengan la misma función.

Observamos que tanto para las proteínas Ploop y NADP no es posible predecir su función de forma satisfactoria, no llegando a obtener un valor ARI ¿0.62, ni un valor VI ¡1 en ninguno de los casos independientemente del algoritmo de agrupamiento utilizado. Además el número óptimo de grupos definidos por los dos índices, k_{VI} y k_{ARI}, pueden distar entre sí en más de 20 grupos y tienden a alejarse de los valores esperados $k_{Ploop} = 53$ y $k_{NADP} = 50$, siendo k_{VI} el que más se acerca a estos valores.

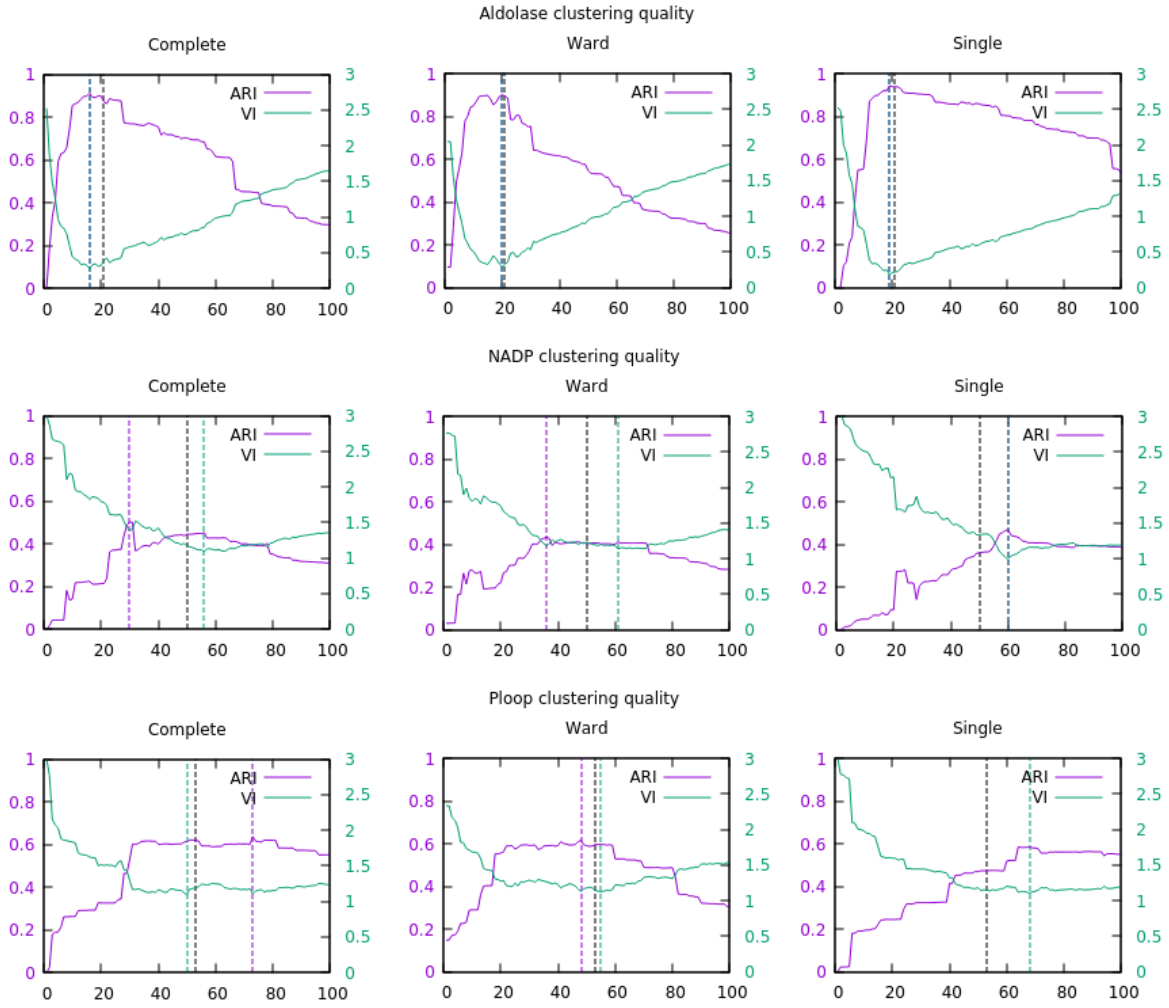

Figura 7: Figura que muestra la evolución de la clasificación con respecto a la clasificación

Podemos decir que los resultados obtenidos muestran que los distintos algoritmos de agrupamiento se comportan de forma distinta, sin embargo la partición optima encontrada con respecto a la clasificación esperada tienen valores parecidos tanto de ARI como de VI. Esto quiere decir que aunque la clasificación optima obtenida por los distintos algoritmos disten mucho entre si cuando la clasificación original no puede ser obtenida, su distancia a la clasificación optima será parecida, quitando importancia al método de agrupamiento jerárquico utilizado.

Por esta razón, de aquí en adelante sólo se usará Ward como método de agrupamiento. En las siguientes secciones nos centraremos en los índices obtenidos a lo largo de este AAAJ, representados en las Figuras 8,9 y 10, analizando los conjuntos de datos por separado.

3.3.2. Aldosasas

En primer lugar observamos que ningún criterio de paro coincide exactamente con la clasificación funcional pero encuentran valores muy cercanos con $ARI = 0{,}91$ y $VI = 0{,}29$ coincidiendo en la partición encontrada, $k_{VI} = k_{ARI} = 20$. Es necesario destacar que $TV = 0$ para $k = 14$ y $k = 15$ por lo que podemos decir que los grupos que conforman esas particiones cumplen la tercera propiedad de una clase de equivalencia y teniendo valores ARI y VI cercanos a los mejores obtenidos.

El punto de paro establecido por CIC, $k = 21$, coincide con un mínimo local de TV y con el número de funciones previamente establecidas, produciendo una partición muy parecida a la correcta con $ARI = 0{,}89$ y $VI = 0{,}32$. Por otro lado los puntos de paro encontrados por SI y CH coinciden en $k = 13$, con valores $ARI = 0{,}89$ y $VI = 0{,}37$, o sea estos criterios confirman su sesgo a juntar grupos más allá de la clasificación de referencia. Las particiones en los punto de paro de estos criterios tienen $TV \neq 0$. El punto de paro establecido para DI no se muestra en la gráfica y es muy cercano al número de elementos ,en el rango mostrado tiene un máximo absoluto en $k = 14$, valor donde $TV = 0$.

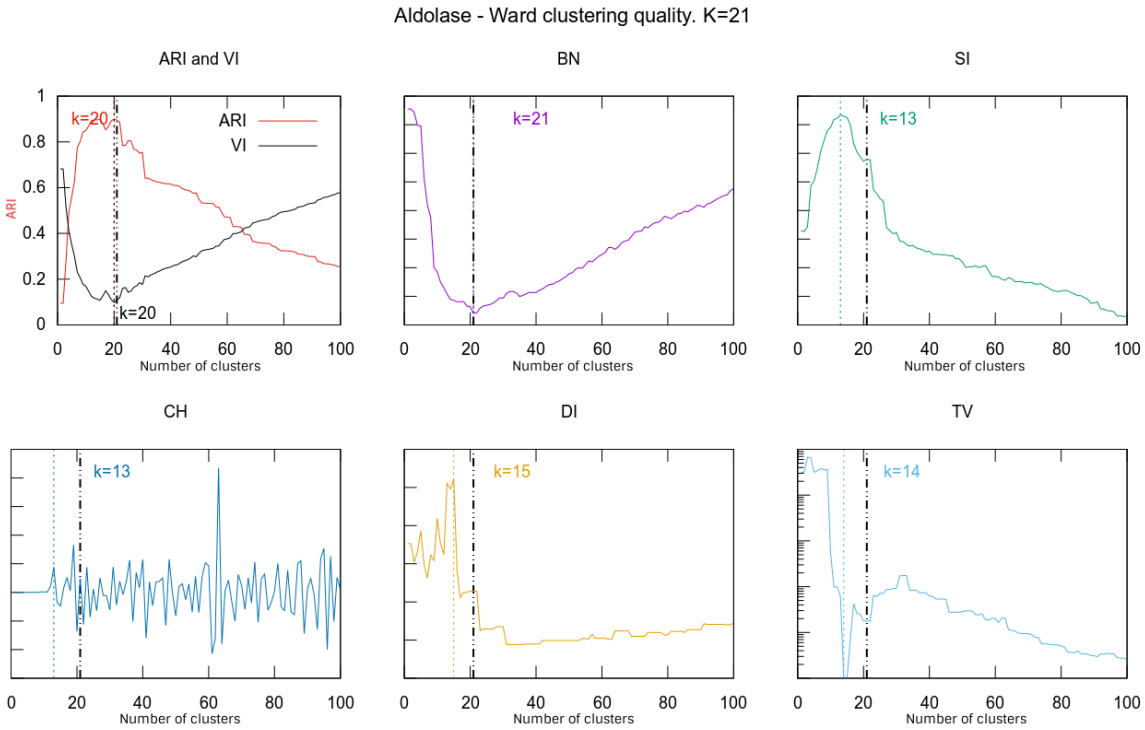

Figura 8: Evolución de los índices a lo largo del algoritmo de Ward aplicado en el conjunto datos de las Aldolasas. Debido a la imposibilidad de representar CH en este rango, se representa $(\partial CH/\partial k)(N-k)$, donde por casualidad coincide el primer máximo local. Se puede comparar el punto de paro, objetivo o no, de cada índice (línea vertical del color del índice) con el número de funciones del conjunto de datos (línea vertical negra)

Con estas observaciones no podemos decir que exista un consenso en el punto de paro

entre los distintos criterios, sin embargo todos ellos encuentran particiones con un valor de similitud aceptable con respecto a la clasificación funcional establecida para este conjunto de dominios. Como era de esperar CH y SI tienden a agrupar más que CIC y DI, aunque ninguno de ellos encuentra las particiones donde la violación de transitividad es nula.

3.3.3. NADP

Como ya hemos comentado en el apartado 3.3.1 en este conjunto de datos no se puede obtener la clasificación funcional esperada. En al Figura 9 CIC muestra un mínimo global (punto de paro) para $k = 69$, y al mismo tiempo tiene varios mínimos locales muy pronunciados tanto en $k = 17$ como en $k = 4$. Esto muestra claramente que para esta medida pueden existir distintos puntos paro para este criterio. El punto de paro encontrado es el más cercano al número de grupos funcionales, se encuentra en un mínimo local de TV y además se encuentra especialmente próximo al mejor valor de VI donde $K_{VI} = 61$.

SI tiene su máximo global en $k = 1$ pero también muestra un máximo de los máximos locales en $k = 60$, lo que, de forma análoga a CIC, abre la posibilidad a establecer criterios de punto de paro distintos respecto al criterio usado en este trabajo.

El primer mínimo local de TV está en $k = 4$ no muestra mínimos locales especialmente pronunciados a largo del AAAJ, pero sí apreciamos que crece de forma escalonada, teniendo el escalón más alto entre $k = 62$ y $k = 63$ teniendo un orden de magnitud de diferencia de esta medida entre ambos puntos.

El criterio establecido con el índice CH establece el punto de paro en $k = 4$ al igual que el establecido por DI si tomamos en cuenta sólo los últimos 100 pasos del AAAJ. De la misma manera CH también presenta un máximo global en este rango en $k = 66$.

3.3.4. Ploop

En este caso, al igual que en caso anterior, no se puede encontrar la clasificación funcional asignada a los dominios de este conjunto de datos. TV crece de forma escalonada a lo largo de algoritmo de agrupamiento, teniendo su primer mínimo local en $K = 5$, y su escalón más pronunciado entre $k = 41$ y $k = 38$ con una diferencia de un orden de magnitud entre ambos valores.

CIC tiene una evolución bastante suave mostrando un claro punto de paro en $k = 17$ coincidiendo con un mínimo local especialmente pronunciado de TV.

Los punto de paro de criterio SI ($k = 5$), CH ($k = 4$) y DI ($k = 4$) indican un valor parecido sin que haya acontecimientos destacables en su evolución.

Todos los puntos de paro están muy alejados del número de optimo de particiones, indicando una tendencia a juntar más que la partición de referencia, o sea puntos de corte muy arriba en el árbol.

4. Discusión

En este trabajo se ha planteado la posibilidad de encontrar un punto de paro automático de AAAJ basados en una matriz de distancia, correspondiente a una partición que agrupe lo más posible sin incurrir en grandes violaciones de la propiedad transitiva por la cual, si dos

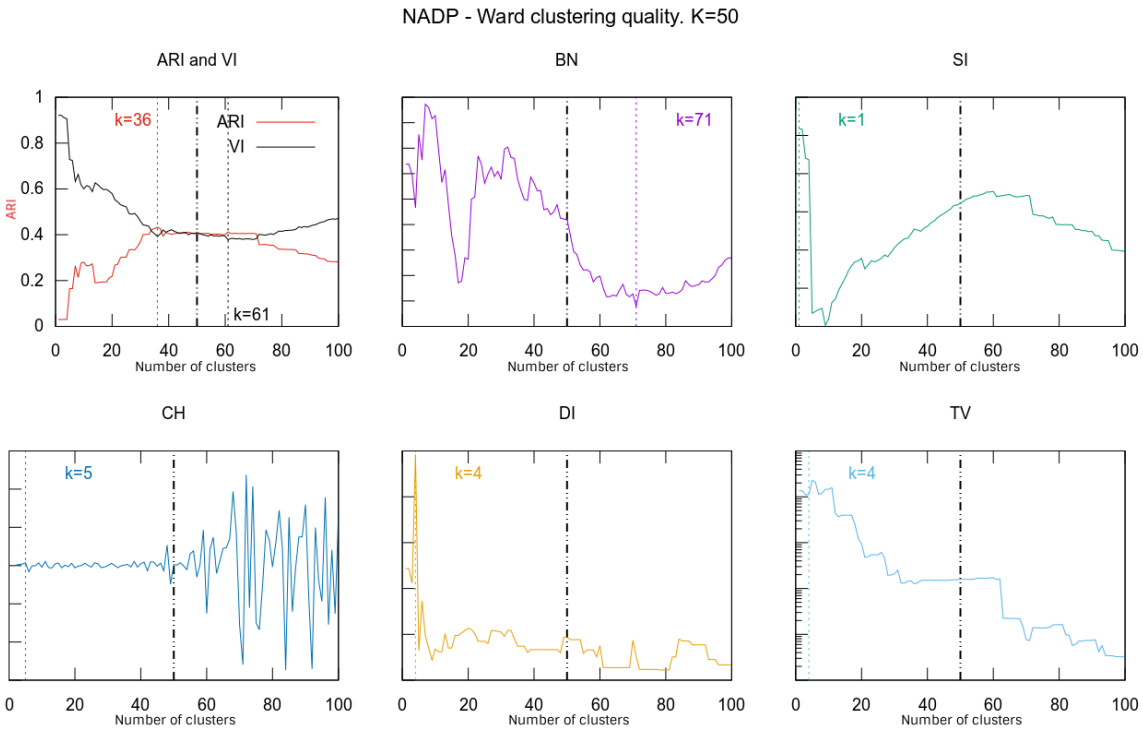

Figura 9: Evolución de los índices a lo largo del algoritmo de Ward aplicado en el conjunto datos NADP. Debido a la imposibilidad de representar CH en este rango, se representa $(\partial CH/\partial k)(N-k)$, donde por casualidad coincide el primer máximo local. Se puede comparar el punto de paro, objetivo o no, de cada índice (línea vertical del color del índice) con el número de funciones del conjunto de datos (línea vertical negra)

elementos están cercanos a un elemento intermedio según un determinado umbral d_0, también están cercanos entre sí según el mismo umbral.

Con este propósito, en primer lugar se ha definido una medida (TV) que cuantifica las violaciones de la propiedad transitiva de las relaciones de equivalencia. Se ha comparado esta medida con una medida análoga ya existentes (BC), basada en el solapamiento de las distancias d^{intra} y d^{inter}. Se ha demostrados que estas medidas están altamente correlacionadas a escala logarítmica, pero TV es más robusta porque se ve mucho menos afectada por el número de elementos que componen los grupos. Además, al no hacer uso de histogramas, TV es menos costosa computacionalmente que BC [7]. Sin embargo no se puede usar TV como criterio de punto de paro porque presenta sus mínimos globales en correspondencia con las particiones triviales $K=1$ y $K=N$.

Por esta razón hemos desarrollado un nuevo criterio (CIC) basado en un índice inspirado en teorá de la información y similar al criterio de información de Akaike (AIC) que nos propone la partición más informativa según determinadas hipótesis, entre las cuales la más importante es que los elementos de cada clase siguen una distribución Gaussiana. CIC ha sido probado y comparado con otros criterios basados en índices existentes (SI,CH y DI). Uno de los índices (SI) ha sufrido modificaciones con respecto al original para poder cumplir con dos objetivos.

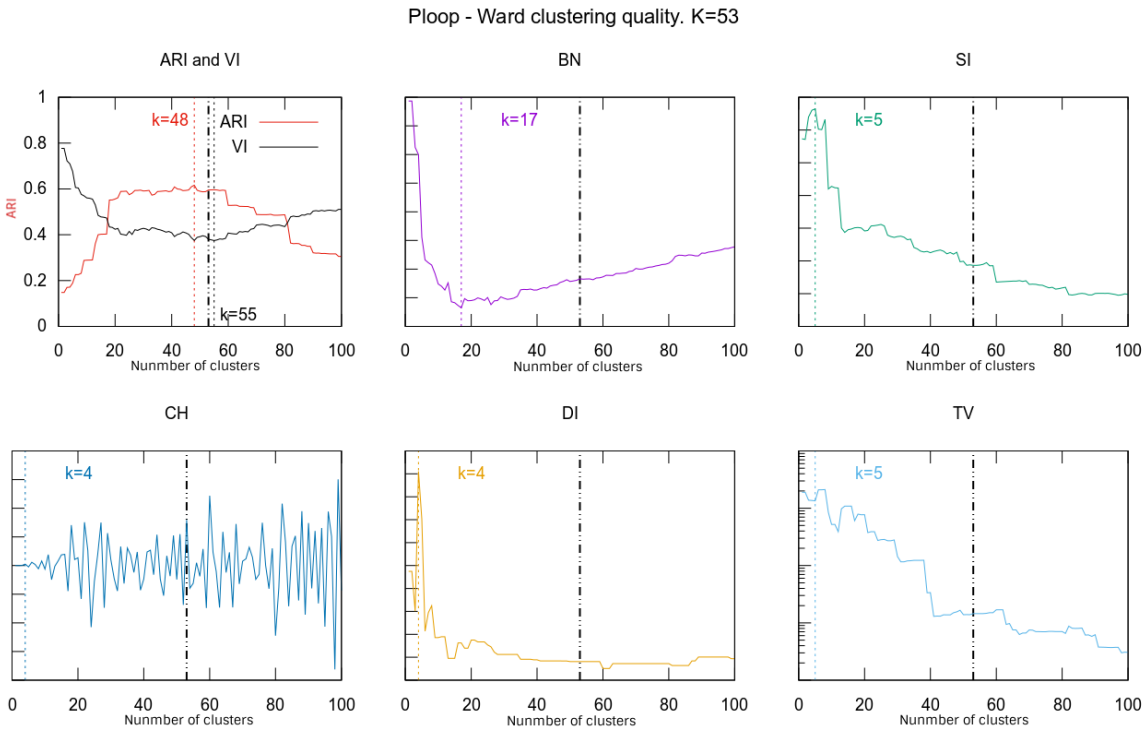

Figura 10: Evolución de los índices a lo largo del algoritmo de Ward aplicado en el conjunto datos NADP. Debido a la imposibilidad de representar CH en este rango, se representa $(\partial CH/\partial k)(N-k)$, donde por casualidad coincide el primer máximo local. Se puede comparar el punto de paro, objetivo o no, de cada índice (línea vertical del color del índice) con el número de funciones del conjunto de datos (línea vertical negra)

Por un lado que nos indique si en un conjunto de datos existen grupos o no, y por el otro que en caso de haberlos permita aplicar un criterio de paro automático. Para el resto de índices (CH y DI) no se han podido hacer las modificaciones necesarias para cumplir esos objetivos, lo que no ha impedido que se hayan podido establecer criterios de punto de paro que han funcionado con éxito.

Para poder caracterizar los índices y evaluar sus criterios de paro, he desarrollado un modelo matemático que genera grupos de puntos en un espacio euclídeo, lo suficientemente semejantes a un conjunto de árboles simétricos, para así comparar la partición obtenida con el criterio de paro con la partición simulada De esta manera, se ha podido comprobar como el número de elementos afecta a los índices, viéndose afectados especialmente en magnitud los índices CIC, CH y DI,. En el caso de DI, esto ha impedido establecer un criterio de paro satisfactorio a lo largo del AAAJ. En estos experimentos el resto de puntos de paro han funcionado de una forma satisfactoria. Cuando la partición original está frustrada, o sea presenta violaciones de transitividad importantes, SI y CH tienden a encontrar puntos de paro cuyo número de grupos es igual o inferior a los originalmente establecidos, y por tanto tienen un sesgo a juntar demasiado. CIC cumple con el propósito para el que fue concebido, estableciendo puntos de corte en el árbol genera bajos cuando la violación de transitividad

entre grupos empieza a aumentar. Cuan do la transitividad aumenta, pasado un umbral, los grupos formados por el AAAJ cuando k es igual al número de grupos originales no definen una distribución gaussiana, y por tanto CIC establece un punto de paro anterior, no permitiendo que se establezcan tales grupos.

Una vez comprobado el comportamiento de los índices, hemos probado tres conjuntos curados de dominios de proteínas (Aldolasas, NADP y Ploop) con la intención de clasificarlos en clases funcionales a partir de una matriz de distancias estructurales. La clasificación resultante se ha comparado con aquella obtenida con los términos de *Gene Ontology*. De todos ellos, sólo en el caso de las Aldolasas la predicción a través de AAAJ es aceptable llegando a un valor de ARI próximo a uno y una valor de VI cercano a cero para todas la metodologías de AAAJ utilizadas. Esto valida, sólo en este caso y de cierta manera, la distancia *contact divergence* utilizada para calcular las distancias por pares de los dominios, aunque los distintos comportamientos de las metodologías utilizadas en los AAAJ (Figura 7) en los tres conjuntos de datos indican que el espacio descrito está lejos de ser ultramétrico. Por que si ese fuera el caso la secuencia de agrupaciones sería la misma, o almenos parecida, para todas la metodologías y por tanto tambien la evolucion de ARI y VI.

Aunque para las Aldolasas los resultados se aproximen al esperado, no se puede decidir de manera objetiva cual de los puntos de paro es el óptimo. Los puntos de paro determinados por SI y CH tienen $TV = 0$, indicando que los grupos de la partición encontrada respetan la propiedad transitiva de las clases de equivalencia. Aunque este punto de paro es razonable, no consigue la máxima información posible porque junta funciones diferentes. El criterio de paro de CIC produce el número correcto de grupos, pero la partición encontrada viola la transitividad y difiere de la partición correcta.

4.1. ¿ Es razonable que exista un único umbral ?

En este trabajo hemos comparado las particiones generadas en cada iteración del AAAJ, que resulta de un corte transversal en el árbol generado (Figura 1) por el algoritmo, sin embargo el conjunto de particiones que se puede obtener parando el algoritmo en cada uno de sus N pasos no contiene la totalidad de los subárboles posibles contenidos en el árbol generado, por tanto no tiene porque tener la partición óptima deseada.

Como muestra la figura 11, cada subárbol está caracterizados por la distancia d^{intra} entre los pares de elementos que lo conforman. Si queremos analizar un conjunto de datos formados por varios árboles definidos por distancias de muy distintas magnitudes puede darse el caso que no podamos aplicar un único punto de paro.

Esto no invalida en ningún caso el árbol generado con el AAAJ, ya que se basa en juntar los elementos que se encuentren a menor distancia independientemente de la metodología utilizada. Sin embargo el número de posibles combinaciones de sub-árboles del árbol generado es mucho mayor que el número de particiones generadas a cada iteración del algoritmo, que además no tiene porque contener la solución optima.

Sin embargo el analizar el árbol con misma metodología planteada hasta el momento de forma iterativa si debería de permitir encontrar la partición optima, al menos para este caso. Esto quiere decir que partiendo del conjunto de particiones generadas en cada paso del algoritmo de agrupamiento encontramos un primer punto de paro que divide sus ramas de nuevo en un conjunto de árboles. Aplicando en la misma metodología por cada conjunto de árboles encontrados para cada punto de paro deberíamos de encontrar en conjunto de árboles

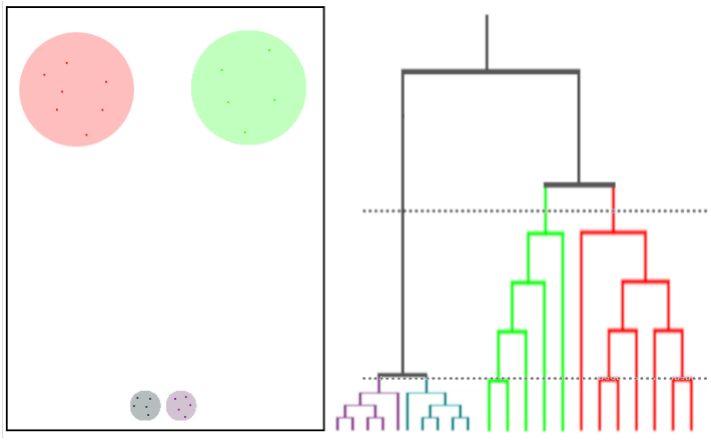

Figura 11: Figura que representa de forma esquemática un conjunto de datos (izquierda) y el árbol (derecha) obtenido como salida de un algoritmo de agrupamiento jerárquico. Se muestra la imposibilidad de encontrar un único punto de corte (lineas discontinuas horizontales) que diferencie los cuatro grupos al que pertenecen originalmente los datos.

que definen la partición optima.

5. Conclusiones

- TV no permite encontrar un punto de paro cuando violación de transitividad entre grupos supera cierto umbral. Además se ha demostrado que esta medida es análoga al solapamiento entre dos funciones de densidad de probabilidad
- CIC es índice que indica claramente cuando los grupos de las particiones formadas a lo largo del AAAJ dejan de seguir una distribución gaussiana.
- Los índices SI y CH permiten establecer un criterio de paro razonable. Teniendo la capacidad de inferir mejor la información subyacente a los datos que DI y CIC.
- En caso de la clasifición de proteínas, los criterios probados no permiten predecir la función de los dominios estableciendo un único punto de paro a lo largo de un AAAJ.

Referencias

[1] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 11, pp. 3747–3752, 2004.

[2] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.

[3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics reports*, vol. 424, no. 4, pp. 175–308, 2006.

[4] L. Bromham and D. Penny, "The modern molecular clock," *Nature Reviews Genetics*, vol. 4, no. 3, p. 216, 2003.

[5] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[6] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *City*, vol. 1, no. 2, p. 1, 2007.

[7] ——, "Comprehensive survey on distance/similarity measures between probability density functions," 2007.

[8] B. V. N. A. Charrad M, Ghazzali N, "Nbclust: An r package for determining the relevant number of clusters in a data set," *ournal of Statistical Software*, vol. 61, no. 6, pp. 1–37, 2014.

[9] C. Chothia and A. M. Lesk, "The relation between the divergence of sequence and structure in proteins." *The EMBO journal*, vol. 5, no. 4, pp. 823–826, 1986.

[10] A. Coordinators: Sebastián and A. Pascual-García, *Bioinformática con Ñ*. Independent Edition, 2015. [Online]. Available: https://es.scribd.com/doc/231270078/Bioinformatica-con-N

[11] S. Das, D. Lee, I. Sillitoe, N. L. Dawson, J. G. Lees, and C. A. Orengo, "Functional classification of cath superfamilies: a domain-based approach for protein function annotation," *Bioinformatics*, vol. 31, no. 21, pp. 3460–3467, 2015.

[12] B. Desgraupes, "Clustering indices," *University of Paris Ouest-Lab Modal'X*, vol. 1, p. 34, 2013.

[13] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," 1973.

[14] L. R. Foulds and R. L. Graham, "The steiner problem in phylogeny is np-complete," *Advances in Applied mathematics*, vol. 3, no. 1, pp. 43–49, 1982.

[15] M. R. Garey and D. S. Johnson, "The rectilinear steiner tree problem is np-complete," *SIAM Journal on Applied Mathematics*, vol. 32, no. 4, pp. 826–834, 1977.

[16] B. Giardina, I. Messana, R. Scatena, and M. Castagnola, "The multiple functions of hemoglobin," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 3, pp. 165–196, 1995.

[17] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[18] R. Kolodny, D. Petrey, and B. Honig, "Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction," *Current opinion in structural biology*, vol. 16, no. 3, pp. 393–398, 2006.

[19] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.

[20] S. Nepomnyachiy, N. Ben-Tal, and R. Kolodny, "Global view of the protein universe," *Proceedings of the National Academy of Sciences*, vol. 111, no. 32, pp. 11 691–11 696, 2014.

[21] A. Pascual-García, D. Abia, R. Méndez, G. S. Nido, and U. Bastolla, "Quantifying the evolutionary divergence of protein structures: the role of function change and function conservation," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 1, pp. 181–196, 2010.

[22] A. Pascual-García, D. Abia, Á. R. Ortiz, and U. Bastolla, "Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures," *PLoS Comput Biol*, vol. 5, no. 3, p. e1000331, 2009.

[23] W. H. Press, *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[24] R. Rammal, G. Toulouse, and M. A. Virasoro, "Ultrametricity for physicists," *Reviews of Modern Physics*, vol. 58, no. 3, p. 765, 1986.

[25] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[26] R. I. Sadreyev, B.-H. Kim, and N. V. Grishin, "Discrete–continuous duality of protein structure space," *Current opinion in structural biology*, vol. 19, no. 3, pp. 321–328, 2009.

[27] W. R. Taylor, "Evolutionary transitions in protein fold space," *Current opinion in structural biology*, vol. 17, no. 3, pp. 354–361, 2007.

[28] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *Journal of Machine Learning Research*, vol. 11, no. Oct, pp. 2837–2854, 2010.

[29] E. Zuckerkandl and L. Pauling, *Molecular disease, evolution and genetic heterogeneity*. Academic Press, 1962, pp. 189–225.